

# BAB I

## PENDAHULUAN

### A. Latar Belakang

Form digital sering digunakan untuk mengumpulkan data, namun sering kali terdapat masalah data yang tidak terisi, yang dapat mempengaruhi kualitas analisis. Untuk mengatasi hal ini, dibutuhkan solusi yang efisien seperti validasi otomatis dengan menggunakan teknologi machine learning. Di dalam penelitian yang berjudul "Penerapan machine learning pada analisis sentimen aplikasi mytelkomsel menggunakan data ulasan google playstore" bahwa machine learning terbukti efektif dalam menganalisis data ulasan aplikasi secara otomatis, dengan algoritma yang mampu mengklasifikasikan sentimen pengguna. Pendekatan serupa dapat diterapkan untuk validasi data kosong pada Form digital, di mana machine learning dapat digunakan untuk mendeteksi dan mengisi data yang hilang, sehingga meningkatkan efisiensi dan kualitas pengolahan data (Fauzan et al., 2024).

Dalam pengembangan model pembelajaran mesin (machine learning), kualitas data yang digunakan memegang peranan penting dalam menentukan kinerja model tersebut. Data yang tidak valid atau rusak dapat mengakibatkan model menghasilkan prediksi yang tidak akurat atau bahkan gagal berfungsi. Oleh karena itu, validasi data menjadi elemen krusial untuk memastikan keandalan dan efektivitas model. Salah satu solusi untuk masalah ini adalah penerapan validasi data secara otomatis dan presisi. Di dalam penelitian yang berjudul "*Automatic and Precise Data Validation for Machine Learning*" mengusulkan pendekatan inovatif dengan memanfaatkan *Partition Summarization (PS)* untuk mendeteksi dan mengatasi masalah data pada berbagai tahap dalam alur kerja pembelajaran mesin. Dengan menggunakan sistem ini, data dapat divalidasi secara otomatis dan akurat, memungkinkan deteksi masalah data lebih awal, serta meningkatkan kualitas dan kinerja model pembelajaran mesin secara keseluruhan (Shankar et al., 2023).

Dengan berkembangnya transformasi digital, manajemen proyek perangkat lunak kini lebih mengarah pada pendekatan Agile yang memungkinkan tim untuk beradaptasi dan merespons perubahan dengan cepat. Meskipun Agile menawarkan fleksibilitas dan meningkatkan kolaborasi, tantangan dalam estimasi upaya, biaya, dan risiko masih ada, yang dapat memengaruhi keberhasilan proyek. Salah satu solusi untuk mengatasi masalah ini adalah penerapan Machine Learning (ML), yang dapat meningkatkan

akurasi prediksi dan efisiensi dalam manajemen proyek. Dengan kemampuannya untuk menganalisis data historis proyek, ML dapat menemukan pola yang tidak terlihat oleh metode konvensional, memungkinkan estimasi yang lebih tepat dan pengelolaan risiko yang lebih baik. Penelitian berjudul *“A systematic literature review on machine learning applications for agile project management”* melakukan tinjauan literatur sistematis untuk menilai penerapan ML dalam manajemen proyek Agile. Studi ini bertujuan mengidentifikasi teknik-teknik ML yang paling umum digunakan serta area yang memiliki potensi untuk dikembangkan lebih lanjut, memberikan wawasan yang berguna bagi praktisi dan peneliti dalam meningkatkan penerapan ML dalam manajemen proyek Agile (Castillo et al., 2024)

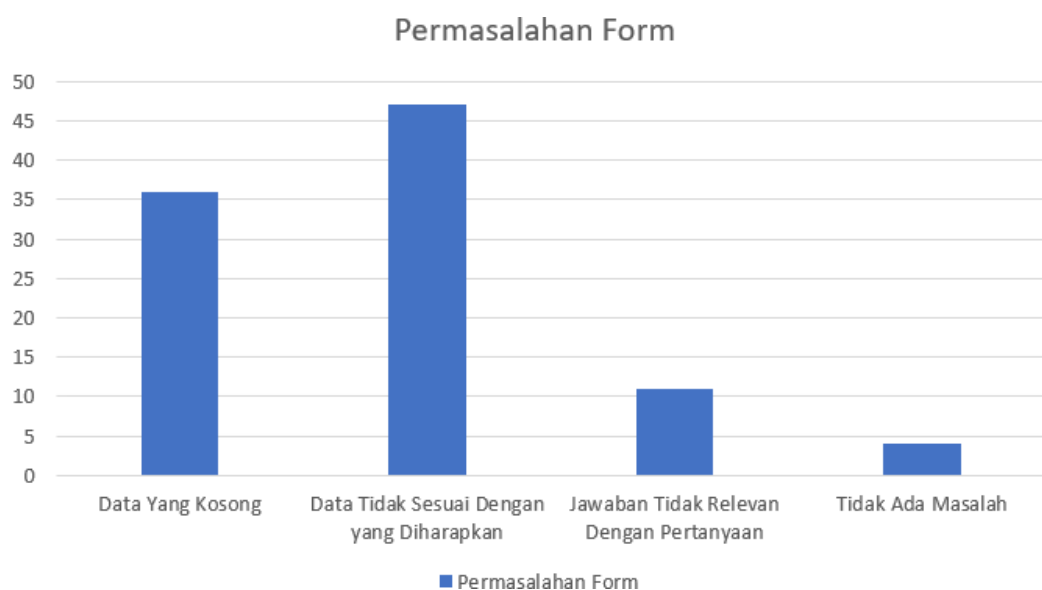
Pengisian data kategorikal dalam formulir sering kali menghadapi tantangan karena rentan terhadap kesalahan manusia, yang dapat mempengaruhi keabsahan data dan ketepatan analisis. Untuk mengatasi permasalahan ini, pendekatan berbasis machine learning telah dikembangkan untuk mengotomatisasi proses pengisian data dengan lebih efisien. Penelitian yang berjudul *“A Machine Learning Approach for Automated Filling of Categorical Fields in Data Entry Forms”* menunjukkan bahwa algoritma machine learning dapat mempelajari pola dari data historis untuk memprediksi nilai kategorikal secara otomatis, sehingga dapat meningkatkan efisiensi, akurasi pengisian data, dan mengurangi ketergantungan pada input manual (Belgacem et al., 2022)

Proses validasi data string memiliki peran penting dalam memastikan data yang dimasukkan sesuai dengan format yang diharapkan. Namun, metode manual yang sering diterapkan kurang efisien, terutama saat harus memproses dataset yang besar dan kompleks. Kesalahan dalam validasi dapat berakibat pada ketidakakuratan data, yang pada akhirnya memengaruhi analisis dan keputusan yang diambil. Penelitian *“Automatic String Data Validation with Pattern Discovery”* menawarkan solusi berupa metode otomatis dengan menggunakan algoritma penemuan pola untuk mengenali pola dalam data string secara lebih akurat. Metode ini membuat proses validasi data menjadi lebih cepat dan efisien, sekaligus memastikan kualitas data yang lebih baik untuk berbagai kebutuhan, seperti pengolahan data bisnis dan pengembangan perangkat lunak (Lin et al., 2024)

Penggunaan formulir online, seperti Form digital, semakin populer di berbagai bidang, termasuk pendidikan, penelitian, dan administrasi. Meskipun formulir ini mudah diakses, pengisian dan pengelolaan data yang terkumpul sering kali memerlukan banyak waktu dan tenaga, terutama jika jumlah pengisian formulir sangat besar. Untuk mengatasi masalah ini, otomatisasi pengisian formulir menjadi solusi yang efektif. Penelitian yang berjudul "*An Application to Automate the Form digital Submission*" mengusulkan pengembangan aplikasi yang dapat mengotomatisasi proses pengisian Form digital, yang dapat meningkatkan efisiensi serta mengurangi kemungkinan terjadinya kesalahan manusia. Aplikasi ini memudahkan pengelolaan data dan mempercepat pengumpulan informasi dengan cara yang lebih akurat dan efisien, yang sangat berguna dalam berbagai situasi pengumpulan data (Kaur et al., 2022)

## B. Permasalahan

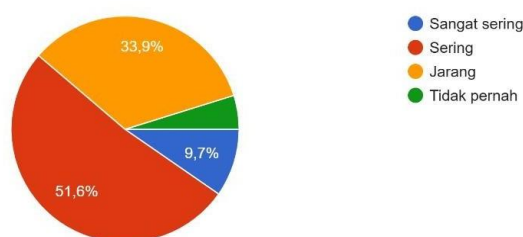
Dalam era digital, penggunaan formulir online atau formulir digital telah menjadi salah satu sarana utama dalam mengumpulkan data, baik di bidang pendidikan, penelitian, maupun administrasi organisasi. Meskipun memberikan kemudahan dalam distribusi dan pengisian, formulir digital masih menyisakan sejumlah permasalahan yang memengaruhi kualitas data yang diperoleh.



**Gambar 1. 1 Permasalahan yang dihadapi**

Berdasarkan hasil survei terhadap 62 responden, dapat diketahui bahwa terdapat empat kategori utama permasalahan yang sering terjadi pada proses pengisian form digital. Permasalahan yang paling dominan adalah data tidak sesuai dengan yang diharapkan, yang mencapai sekitar 47% dari total responden. Hal ini menunjukkan bahwa hampir setengah dari pengguna mengisi form dengan data yang tidak relevan atau tidak sesuai format yang diharapkan, seperti kesalahan penulisan atau pilihan jawaban yang tidak tepat. Selanjutnya, permasalahan data yang kosong menempati posisi kedua dengan persentase sekitar 36%. Kondisi ini menggambarkan bahwa masih banyak pengguna yang tidak mengisi seluruh kolom form, sehingga menyebabkan kekosongan data yang dapat menghambat proses analisis dan validasi. Kemudian, jawaban yang tidak relevan dengan pertanyaan mencatat angka sekitar 11%, menunjukkan bahwa sebagian kecil responden memberikan jawaban yang tidak sesuai konteks. Sementara itu, hanya sekitar 6% responden yang menyatakan tidak mengalami masalah dalam pengisian form. Dari hasil tersebut dapat disimpulkan bahwa mayoritas pengguna masih menghadapi kendala dalam pengisian form, terutama pada aspek ketidaksesuaian dan kekosongan data. Oleh karena itu, dibutuhkan solusi berupa sistem validasi otomatis, misalnya melalui penerapan algoritma *machine learning*, agar data yang diinput lebih akurat dan sesuai dengan ketentuan yang diharapkan.

Seberapa sering Anda menghadapi data kosong atau data yang tidak sesuai format?  
62 jawaban



**Gambar 1. 2 Permasalahan User**

Jika ditinjau dari frekuensi, 51,6% responden sering menghadapi permasalahan data kosong atau tidak sesuai format, bahkan 33,9% responden menyatakan sangat sering mengalaminya. Hanya sebagian kecil responden, yakni 9,7%, yang jarang mengalami kendala tersebut, dan 4,8% lainnya tidak pernah mengalaminya. Fakta ini mengindikasikan bahwa masalah validitas data merupakan isu umum yang dialami pengguna formulir digital.

Permasalahan data kosong dan ketidaksesuaian format tersebut tidak hanya memperlambat proses analisis, tetapi juga berisiko menurunkan akurasi dalam pengambilan keputusan. Dengan semakin meningkatnya ketergantungan pada data

digital, maka dibutuhkan sebuah solusi yang mampu melakukan validasi otomatis terhadap data kosong atau tidak valid pada formulir online, sehingga kualitas data tetap terjaga dan efisiensi pengolahan informasi dapat tercapai.

## **1. Identifikasi Masalah**

Berdasarkan permasalahan hambatan yang di hadapi dapat diidentifikasi masalahnya adalah :

- a. Banyak responden melaporkan adanya data kosong dalam hasil pengumpulan melalui form digital, yang dapat mengurangi kualitas dan keandalan data yang diperoleh.
- b. Sebagian besar pengguna mengalami ketidaksesuaian antara data yang diterima dengan harapan mereka, yang menunjukkan adanya masalah dalam desain pertanyaan atau pemahaman responden terhadap pertanyaan yang diajukan.

## **2. Rumusan Masalah**

### **a. Research Questions**

Pertanyaan penelitian yang dirumuskan adalah sebagai berikut:

- a) Bagaimana metode *K-Nearest Neighbors (KNN)* dapat digunakan untuk mendeteksi dan memvalidasi data kosong secara otomatis pada form digital?
- b) Sejauh mana metode *K-Nearest Neighbors (KNN)* mampu meningkatkan akurasi dalam proses validasi data dibandingkan dengan metode konvensional?

### **b. Problem Statement**

Saat ini menunjukkan peningkatan penggunaan alat digital seperti Form digital untuk pengumpulan data di berbagai sektor, namun tantangan signifikan muncul terkait ketidaklengkapan dan ketidakakuratan data yang sering dihadapi pengguna. Banyak responden tidak memberikan jawaban yang lengkap atau relevan, yang dapat mengurangi kualitas data dan memengaruhi analisis serta pengambilan keputusan. Selain itu, keterbatasan dalam penerapan sistem verifikasi otomatis dan ketergantungan pada proses manual meningkatkan risiko kesalahan. Oleh karena itu, ada kebutuhan mendesak untuk mengembangkan solusi yang lebih efektif dan otomatis dalam pengelolaan data guna memastikan keandalan informasi yang digunakan.

## C. Maksud dan Tujuan Penelitian

### 1. Maksud Penelitian

Maksud dari penelitian ini adalah menerapkan *machine learning* dengan algoritma *K-Nearest Neighbors* (KNN) untuk melakukan validasi data pada platform pengumpulan data seperti Form digital. Sistem yang dibangun bertujuan mendeteksi entri kosong secara otomatis berdasarkan pola data historis, sehingga proses validasi menjadi lebih cepat, akurat, dan mengurangi kebutuhan pengecekan manual.

### 2. Tujuan Penelitian

Penelitian ini bertujuan mengembangkan metode verifikasi dan validasi data berbasis *machine learning* menggunakan algoritma *K-Nearest Neighbors* (KNN) pada platform pengumpulan data form digital. Tujuannya adalah memastikan setiap data yang masuk lengkap dan valid secara otomatis, sehingga mengurangi kesalahan input dan meningkatkan kualitas Data yang dihasilkan.

2 fokus utama dalam pengembangan penelitian :

- a. Mengembangkan metode verifikasi dan validasi data otomatis berbasis machine learning untuk meningkatkan kualitas data yang dikumpulkan melalui form digital.
- b. Memberikan efektivitas untuk mengatasi permasalahan data kosong dalam pengelolaan data.

## D. Spesifikasi Hasil

### 1. Sistem Verifikasi Data Otomatis

- a. Secara otomatis mendeteksi, memperbaiki, dan mencegah adanya data kosong atau tidak valid pada platform seperti Form digital.
- b. Memberikan umpan balik langsung kepada pengguna dalam waktu nyata.

### 2. Peningkatan Kualitas Data

Data menjadi lebih lengkap dan valid, sehingga mendukung analisis dan pengambilan keputusan yang lebih tepat.

### 3. Efisiensi Pengolahan Data

Proses validasi dilakukan secara otomatis, menghemat waktu dan tenaga dibandingkan metode manual.

### 4. Pendekatan Gabungan Rule-Based dan Machine Learning

Memadukan aturan logis (rule-based) dan teknologi pembelajaran mesin untuk mendeteksi serta memperbaiki kesalahan data.

### **5. Framework Pengembangan Fleksibel**

Sistem dibangun dengan pendekatan Agile, mempermudah penyesuaian terhadap kebutuhan pengguna yang terus berubah.

### **6. Visualisasi Hasil Validasi**

Menyediakan laporan berbentuk visual, seperti grafik, untuk mempermudah analisis dan penyampaian hasil validasi data.

### **7. Panduan dan Dokumentasi Penggunaan**

Menyediakan dokumentasi yang jelas dan mudah diikuti untuk instalasi, konfigurasi, dan penggunaan sistem, yang dapat diterapkan dalam berbagai situasi.

## **E. Signifikansi Penelitian & Pengembangan/Kegunaan Pembahasan**

Hasil yang diharapkan dari penelitian ini memiliki signifikansi penelitian yang kuat dan konteks penerapan Machine Learning dengan metode Agile yang dimana saat melakukan pengembangan pada perangkat lunak Website dapat memiliki dampak positif yang substansial terhadap perbaikan proses dan memberikan nilai tambah pada praktik industri dan ilmu pengetahuan perangkat lunak secara keseluruhan.

### **1. Manfaat Praktis**

#### **a. Efisiensi dalam Pengolahan Data:**

Sistem ini dirancang untuk profesional dan praktisi, seperti manajer data, pengembang sistem, atau operator platform online, yang memerlukan solusi efisien untuk memastikan keakuratan data.

#### **b. Aplikasi yang Luas:**

Memberikan solusi yang bermanfaat bagi organisasi, perusahaan, atau institusi yang sering menangani volume data besar, guna mengurangi kesalahan dan meningkatkan ketepatan data.

#### **c. Peningkatan Keandalan Data:**

Memastikan pengguna platform online, seperti Form digital, dapat mengumpulkan data yang lebih lengkap dan valid.

## 2. Manfaat Teoritis

### a. Pengembangan Model Gabungan:

Menjadi dasar bagi peneliti dan akademisi di bidang teknologi informasi dan analitik data untuk mengembangkan metode validasi yang lebih canggih.

### b. Referensi untuk Penelitian:

Memberikan referensi yang berguna bagi mahasiswa atau dosen yang tengah melakukan penelitian terkait sistem berbasis aturan dan *machine learning*.

### c. Pendalaman dalam Teknologi Validasi Data:

Membantu komunitas ilmiah untuk lebih memahami efektivitas kombinasi pendekatan berbasis aturan dan *machine learning* dalam meningkatkan kualitas data.

## 3. Manfaat Kebijakan

### a. Mendukung Digitalisasi:

Berguna bagi pemerintah, lembaga publik, atau pengambil kebijakan yang ingin mendorong penerapan teknologi dalam pengelolaan data.

### b. Peningkatan Standar Kualitas Data:

Membantu organisasi atau lembaga standar, seperti ISO atau badan pengawas data, untuk menetapkan pedoman baru dalam pengumpulan dan validasi data.

### c. Pengambilan Keputusan Berdasarkan Data:

Menyediakan dukungan bagi pemimpin organisasi atau pembuat kebijakan untuk membuat keputusan berbasis data yang lebih akurat dan dapat diandalkan.

## F. Asumsi dan Keterbatasan

Asumsi ini didasarkan pada penelitian sebelumnya bahwa data historis memiliki kualitas yang baik untuk melatih algoritma machine learning dalam mendeteksi dan memperbaiki data kosong. Metode Agile digunakan untuk memastikan pengembangan berjalan iteratif dan fleksibel, memungkinkan penyesuaian sesuai kebutuhan dan menghasilkan solusi yang efisien serta berkualitas tinggi.



## **1. Asumsi**

### **a. Keandalan Platform Pengumpulan Data**

Diasumsikan bahwa data yang diperoleh melalui Form digital atau platform serupa dapat diakses dan diekspor tanpa adanya masalah teknis.

### **b. Ketersediaan Data yang Memadai**

Dataset yang digunakan dalam pengujian sistem diharapkan memiliki variasi yang cukup untuk melatih algoritma *machine learning* dan menguji pendekatan berbasis aturan (*rule-based*).

### **c. Kesesuaian dengan Berbagai Konteks**

Pendekatan yang dikembangkan diharapkan dapat diterapkan secara relevan di berbagai sektor, meskipun sistem diuji dalam konteks tertentu.

## **2. Keterbatasan**

### **a. Skalabilitas Sistem**

Sistem mungkin mengalami penurunan kinerja ketika diterapkan pada data dalam jumlah besar karena pengolahan data tidak didukung oleh optimasi dari basis data relasional seperti MySQL.

### **b. Ketergantungan pada Penyimpanan Alternatif**

Tanpa menggunakan MySQL, sistem bergantung pada penyimpanan alternatif seperti file JSON, CSV, atau NoSQL, yang dapat membatasi kemampuan dalam melakukan operasi kompleks seperti pencarian data atau analisis multidimensi. Penyimpanan data tanpa MySQL dapat menurunkan efisiensi, terutama saat pengolahan data yang memerlukan proses seperti pengurutan, penggabungan, atau penyaringan secara cepat.

### **c. Penggunaan Algoritma K-Nearest Neighbors**

Keterbatasan algoritma K-Nearest Neighbors (KNN) terletak pada ketergantungannya terhadap jumlah data dan dimensi fitur yang tinggi, karena semakin banyak data yang digunakan, waktu komputasi dan kebutuhan memori akan meningkat secara signifikan. Selain itu, KNN sangat sensitif terhadap data yang tidak relevan atau fitur dengan skala yang berbeda, sehingga memerlukan proses normalisasi dan seleksi fitur yang baik agar hasilnya akurat. Algoritma ini juga tidak bekerja optimal pada data yang mengandung noise atau outlier, karena perhitungan jarak dapat mudah terpengaruh oleh data yang salah. Terakhir, KNN tidak memiliki proses pembelajaran eksplisit, sehingga setiap kali dilakukan prediksi, perhitungan jarak harus dilakukan ulang terhadap seluruh data latih.

## **E. Definisi Istilah dan Definisi Operasional**

### **1. Machine Learning (Pembelajaran Mesin)**

Machine learning adalah cabang kecerdasan buatan yang memungkinkan sistem untuk belajar dari data tanpa perlu pemrograman yang eksplisit. Dalam penelitian ini, machine learning digunakan untuk mengidentifikasi pola dalam data dan melakukan proses validasi secara otomatis.

### **2. Rule-Based System (Sistem Berbasis Aturan)**

Sistem berbasis aturan adalah metode pengolahan data yang bergantung pada serangkaian aturan logis untuk membuat keputusan atau menyelesaikan masalah. Dalam penelitian ini, pendekatan ini digunakan untuk memverifikasi kelengkapan dan keabsahan data yang dikumpulkan.

### **3. Agile Methodology (Metodologi Agile)**

Agile adalah pendekatan pengembangan perangkat lunak yang menekankan kerja tim yang kolaboratif, siklus pengembangan yang cepat, dan kemampuan untuk menyesuaikan diri dengan perubahan. Dalam konteks penelitian ini, metode Agile diterapkan untuk merancang sistem yang dapat dengan mudah diubah sesuai kebutuhan pengguna.

### **4. Validasi Data (Data Validation)**

Validasi data adalah prosedur untuk memastikan data yang dikumpulkan sesuai dengan format atau kriteria yang telah ditentukan. Proses ini sangat penting untuk menjamin bahwa data yang digunakan untuk analisis atau pengambilan keputusan adalah akurat dan berkualitas.

### **5. Verifikasi Data (Data Verification)**

Verifikasi data adalah tahap untuk memastikan data yang dikumpulkan adalah benar, lengkap, dan konsisten. Dalam penelitian ini, verifikasi dilakukan untuk mencegah adanya kesalahan data, seperti data yang hilang atau tidak valid, yang dapat memengaruhi hasil penelitian.

### **6. NoSQL Database**

NoSQL adalah jenis sistem basis data yang dirancang untuk mengelola data besar dan tidak terstruktur, berbeda dengan sistem database relasional yang tradisional. NoSQL cocok digunakan untuk pengolahan data besar yang tidak memerlukan struktur tabel tetap, seperti dalam format file JSON atau CSV.